Original Paper

Long COVID Discourse in Canada, the United States, and Europe: Topic Modeling and Sentiment Analysis of Twitter Data

Ahmed Ghassan Tawfiq AbuRaed[1], PhD; Emil Azuma Prikryl[2], MD; Giuseppe Carenini[1], Prof Dr; Naveed Zafar Janjua[3], DrPH, Prof Dr Med

[1]Department of Computer Science, The University of British Colombia, Vancouver, BC, Canada

[2]NOSM University, Thunder Bay, BC, Canada

[3]BC Centre for Disease Control, Vancouver, BC, Canada

Corresponding Author:
Ahmed Ghassan Tawfiq AbuRaed, PhD
Department of Computer Science
The University of British Colombia
2366 Main Mall
ICICS Building
Vancouver, BC, V6T 1Z4
Canada
Phone: 1 7789294413
Email: ahmed.aburaed@ubc.ca

Abstract

Background: Social media serves as a vast repository of data, offering insights into public perceptions and emotions surrounding significant societal issues. Amid the COVID-19 pandemic, long COVID (formally known as post–COVID-19 condition) has emerged as a chronic health condition, profoundly impacting numerous lives and livelihoods. Given the dynamic nature of long COVID and our evolving understanding of it, effectively capturing people's sentiments and perceptions through social media becomes increasingly crucial. By harnessing the wealth of data available on social platforms, we can better track the evolving narrative surrounding long COVID and the collective efforts to address this pressing issue.

Objective: This study aimed to investigate people's perceptions and sentiments around long COVID in Canada, the United States, and Europe, by analyzing English-language tweets from these regions using advanced topic modeling and sentiment analysis techniques. Understanding regional differences in public discourse can inform tailored public health strategies.

Methods: We analyzed long COVID–related tweets from 2021. Contextualized topic modeling was used to capture word meanings in context, providing coherent and semantically meaningful topics. Sentiment analysis was conducted in a zero-shot manner using Llama 2, a large language model, to classify tweets into positive, negative, or neutral sentiments. The results were interpreted in collaboration with public health experts, comparing the timelines of topics discussed across the 3 regions. This dual approach enabled a comprehensive understanding of the public discourse surrounding long COVID. We used metrics such as *normalized pointwise mutual information* for coherence and *topic diversity* for diversity to ensure robust topic modeling results.

Results: Topic modeling identified five main topics: (1) long COVID in people including children in the context of vaccination, (2) duration and suffering associated with long COVID, (3) persistent symptoms of long COVID, (4) the need for research on long COVID treatment, and (5) measuring long COVID symptoms. Significant concern was noted across all regions about the duration and suffering associated with long COVID, along with consistent discussions on persistent symptoms and calls for more research and better treatments. In particular, the topic of persistent symptoms was highly prevalent, reflecting ongoing challenges faced by individuals with long COVID. Sentiment analysis showed a mix of positive and negative sentiments, fluctuating with significant events and news related to long COVID.

Conclusions: Our study combines natural language processing techniques, including contextualized topic modeling and sentiment analysis, along with domain expert input, to provide detailed insights into public health monitoring and intervention. These findings highlight the importance of tracking public discourse on long COVID to inform public health strategies, address misinformation, and provide support to affected individuals. The use of social media analysis in understanding public health issues is underscored, emphasizing the role of emerging technologies in enhancing public health responses.

(J Med Internet Res 2024;26:e59425) doi: 10.2196/59425

KEYWORDS

long COVID; topic modeling; sentiment analysis; Twitter; public perception; social media analysis; public health

Introduction

As of December 2023, there have been more than 700 million cases of COVID-19 globally, leading to nearly 7 million deaths [1]. These figures likely constitute an underestimation, given reduced reporting requirements for COVID-19 in most countries.

Beyond the acute effects of infection, a significant portion of survivors of COVID-19 experience a broad spectrum of ongoing symptoms several months after infection, which are generally captured under terms such as long COVID and post–COVID-19 condition. The estimated proportion of survivors with long COVID varies extremely widely from below 10% to around 60% across different studies [2], but population-based studies have reported that about 20% of people developed long COVID [3].

While the definition of long COVID continues to evolve, it has been defined as new, returning, or ongoing health problems that persist or occur 4 or more weeks after SARS-CoV-2 infection, with wide-ranging symptoms that can include fatigue, postexertional malaise, shortness of breath, palpitations, trouble sleeping, cognitive deficits, anxiety, and depression [4].

The characteristics and subtypes of long COVID, which likely represents several overlapping entities, continue to be an active area of research using both traditional observational study designs [5] and other innovative data-driven approaches [6].

During the COVID-19 pandemic, people have used social media such as Twitter (rebranded as "X" in 2023) to share information, opinions, and sentiments about COVID-19 and long COVID. This kind of information can help inform health care organizations and public health organizations and assist them in developing approaches and interventions that are sensitive to the concerns and perceptions of the public. This is of relevance to long COVID because it is a diverse condition that is experienced and perceived by the public in many different ways, making the identification of key issues and concerns associated with the condition of interest to health organizations to inform their communications, interventions, and research objectives around long COVID.

The use of topic modeling and sentiment analysis is widespread for identifying issues and public opinions in the field of public health. These approaches are also being used to gain insights into COVID-19–related matters. Analyses have been performed to unveil patterns in health communications across a variety of data sources, communities, and geographic locations.

While certain studies explored news articles [7] or research papers [8,9], the majority of research concentrated on social media platforms such as Reddit posts [10,11] and tweets [10,12-31]. Various approaches have been used to address the tasks of topic modeling, its visualization, and sentiment analysis.

Latent Dirichlet allocation (LDA) has been widely regarded as the best approach for topic modeling until recent years, due to its probabilistic foundation and effectiveness in uncovering latent themes within a corpus of text [32]. While most of the studies have used LDA to achieve topic modeling [8,12,14,16,20,24,33,34], Ridhwan and Hargreaves [13] have used Gibbs Sampling Dirichlet Multinomial Mixture [35], while Zheng et al [36] used conditional random field for a named entity recognition task aiming to extract the most frequent terms and applying the Jaccard similarity coefficient. Furthermore, Yan et al [37] applied biterm topic modeling, and Sussman et al [10] have used Cision's Brandwatch software to perform both topic modeling and sentiment analysis.

On the other hand, sentiment analysis has been tackled by various rule-based and machine learning methods. Valence Aware Dictionary for Sentiment Reasoning [38] is a famous rule-based system that has been used by many studies [12-14]. TextBlob is a library for processing text that also implements a rule-based sentiment analysis system used by several studies [17,20,22]. Marcec and Likic [18] have used the AFINN lexicon [39], a list of English terms manually rated for valence.

Moreover, other studies have used machine learning methods to improve sentiment analysis. These include classical machine learning methods such as logistic regression, AdaBoost [39], XGBoost [40], and advanced techniques such as long short-term memory networks [41], which have been used by Guo et al [21]. Multilayer perceptron [42], naïve Bayes [43], and support vector machine [44] algorithms have been used by Masood et al [23]. In addition, Kumar et al [19] used bidirectional encoder representations from transformers [45], a deep neural network based on transformer architecture, demonstrating the efficacy of these advanced models in capturing the nuances of sentiment in social media text.

By integrating these approaches, researchers have been able to develop more sophisticated models for sentiment analysis, contributing to a deeper understanding of public opinion and emotional responses on social media platforms. Moreover, some studies have also explored sentiments in the context of specific aspects selected by domain experts [46]. However, such methods have not been used extensively to investigate issues related to long COVID, except for a few select papers (Table 1). While the research involved variations of topic modeling or sentiment analysis techniques, they generally did not combine these approaches.

Table 1. Related work on topic modeling and sentiment analysis on long COVID–related data[a].

Authors	Source	Posters	Time	Location	Language	Topic modeling	Sentiment analysis
Déguilhem et al [47]	Twitter, Reddit, Doctissimo, Facebook, and other forums	Public	January 1, 2020, to August 10, 2021	France	French	Yes, biterm topic modeling [37]	No
Bhattacharyya et al [48]	Twitter	Public	August 28, 2022, to September 6, 2022	Not specified	English	No	Yes, National Research Council Emotion Lexicon [49]
Southwick et al [33]	Reddit	Public	Not specified	Not specified	English	Yes, LDA[b] [32]	Yes, Affective Norms for English Words lexicon [50]
Miyake and Martin [51]	Twitter, Facebook, blogs, news posts on social media, Reddit, forums, and other platforms	Public	January 1, 2020, to January 1, 2021	United Kingdom	English	No	Yes, but only of hashtags and emojis using IBM Watson emotional lexicon
Fu [34]	Twitter	Public	March 26, 2022, to April 26, 2022	United States	English	Yes, LDA [32]	Yes, VADER[c] [38]
Ramakrishnan et al [52]	Twitter	Public	May 1 to Sep 30, 2021	Not specified	English	No	Yes, IBM Watson Tone Analyzer and 6 classical ML[d] algorithms

[a]This table summarizes previous studies on topic modeling and sentiment analysis across various platforms and time frames. It includes information on the source of data, the population being studied, the time period covered, the geographic location, the language used, and the specific methodologies applied for topic modeling and sentiment analysis.

[b]LDA: latent Dirichlet allocation.

[c]VADER: Valence Aware Dictionary for Sentiment Reasoning.

[d]ML: Machine learning.

We have used new methods for our tasks that, according to our knowledge, have never been used before in the context of long COVID analysis. First, we used contextualized topic modeling (CTM) [53], which takes advantage of the entire context of the text rather than LDA, which, by applying neural variational techniques, uses a bag-of-words representation of the text, meaning that it ignores the order of words and considers only the frequency of words. Also, we use Llama 2 [54] to represent the text and classify its sentiment rather than rule-based methods or classical machine learning methods.

In summary, this paper addressed 3 key knowledge gaps related to long COVID and methods for topic modeling and sentiment analysis. First, it investigated topics and sentiments related to long COVID across 3 regions (Canada, the United States, and Europe), providing an opportunity for comparative analysis of issues and public perceptions in different geographical contexts. Second, the study introduced new techniques, such as CTM and large language model (LLM) sentiment analysis, to obtain nuanced insights related to long COVID. These advanced methods surpass traditional approaches by capturing the context and subtleties of the discourse more effectively. Third, the study integrated expertise from public health and computer sciences, combining methodological rigor with domain-specific knowledge to interpret the results in a meaningful way. This interdisciplinary approach ensures that the analytical methods are aligned with public health objectives and implications. By addressing these gaps, the study contributes valuable insights into public health monitoring and intervention strategies, demonstrating the use of social media analysis in understanding complex public health issues such as long COVID.

The innovation in this study lies in the integration of CTM and an LLM (Llama 2) for sentiment analysis. CTM goes beyond traditional LDA by capturing word meanings in context, leading to more coherent and semantically meaningful topics. Using Llama 2 for sentiment analysis provides a nuanced understanding of sentiments expressed in tweets, leveraging the model's robust language comprehension capabilities.

Methods

Data and Data Processing

We used Twitter data for this analysis. In collaboration with public health experts working on long COVID, we identified relevant hashtags and keywords to extract Twitter data related to long COVID. Table 2 provides the list of hashtags and keywords that were used to collect the long COVID–relevant tweets.

Table 2. Terms used to collect long COVID–relevant tweets provided by health care professionals[a].

Term	Description
#Long COVID	This hashtag refers to the long-term effects and symptoms experienced by individuals after recovering from COVID-19.
#mecfs	Abbreviation for "Myalgic Encephalomyelitis/Chronic Fatigue Syndrome," a complex and debilitating condition characterized by extreme fatigue that is not alleviated by rest.
#MyalgicEncephalomyelitis	Also known as ME[b], this is a chronic condition characterized by profound fatigue, pain, sleep disturbances, and other symptoms.
#Fibromyalgia	A chronic disorder characterized by widespread musculoskeletal pain, fatigue, and tenderness in localized areas.
#PostViralSyndrome	A condition that occurs after a viral infection, characterized by persistent symptoms such as fatigue, joint pain, and cognitive issues.
#Dysautonomia	A disorder of the autonomic nervous system, which controls involuntary bodily functions such as heart rate, blood pressure, and digestion. Symptoms can include lightheadedness, fainting, and difficulty regulating body temperature.
#ChronicFatigueSyndrome	Also known as CFS[c], this is a complex disorder characterized by extreme fatigue that does not improve with rest and may be worsened by physical or mental activity.
#PwME	Abbreviation for "People with Myalgic Encephalomyelitis," a term used to refer to individuals living with ME/CFS.
#MyalgicE	An abbreviation for ME, a chronic condition characterized by extreme fatigue and other symptoms.
#PostCovidSyndrome	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.
#postcovid	A term used to describe symptoms or conditions that persist after recovering from COVID-19.
#PostViralFatigueSyndrome	A condition characterized by persistent fatigue and other symptoms following a viral infection.
#postviralillness	A term used to describe the lingering effects of a viral illness, which can include fatigue, muscle pain, and cognitive difficulties.
"Long Covid," "long COVID," and "long COVID syndrome"	A term used to describe the long-term effects of COVID-19, including persistent symptoms such as fatigue, shortness of breath, and cognitive issues.
"Post acute COVID"	A term used to describe the period of time after a COVID-19 infection where symptoms continue to persist.
"Post-COVID syndrome"	A term used to describe the syndrome of lingering symptoms and health issues experienced by individuals after recovering from COVID-19.
"Post-acute sequelae of SARS-CoV-2"	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.
"Long-term COVID"	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.
"Long haulers"	A colloquial term used to describe individuals who continue to experience symptoms and health issues after recovering from COVID-19.
"Chronic COVID syndrome"	A term used to describe the ongoing symptoms and health issues experienced by individuals after recovering from COVID-19.

[a] This table lists the hashtags and keywords identified by public health experts to extract tweets related to long COVID from Twitter. Each term includes a description to explain its relevance to long COVID.

[b] ME: myalgic encephalomyelitis.

[c] CFS: chronic fatigue syndrome.

We focused on English-language tweets due to the availability and accessibility of large volumes of Twitter data in English, allowing for more accurate and reliable analysis. Moreover, this focus enables a comparative study across regions where English is the predominant language, reducing the complexity of multilingual analysis and ensuring the consistency of sentiment and topic modeling methods.

The hashtags in Table 2 were selected based on their relevance to long COVID as identified by public health experts. These experts provided terms frequently used in discussions related to long COVID, ensuring comprehensive data collection. However, we acknowledge that some terms such as #Fibromyalgia, #Dysautonomia, and #ChronicFatigueSyndrome may not exclusively refer to long COVID. To mitigate this, we applied additional filtering and manual inspection to confirm that the majority of tweets were indeed related to long COVID. Despite these efforts, we recognize that some tweets may not be entirely relevant, and this limitation is addressed in the *Discussion* section.

For this study, we collected tweets alongside their metadata for the year 2021. We used the Twitter application programming interface [55] to extract the data, and afterward, we applied necessary preprocessing of the tweets. For preprocessing, we used the *tweet-preprocessor* toolkit, which applies cleaning, tokenizing, and parsing of URLs, hashtags, mentions, reserved words ("RT," which stands for "Retweet" and "FAV," which stands for "Favorite"), emojis, and smileys.

Among the 814,951 tweets extracted in total (Canada: n=98,796, United States: n=289,856, and Europe: n=426,299), we included only those tweets written in English using tweet metadata and the *spacy-langdetect* toolkit [56]. This process resulted in 782,089 tweets in total: 95,743 for Canada, 278,392 for the United States, and 407,954 for Europe including the United Kingdom. To remove tweet-specific keywords and URLs, we used the *tweet-preprocessor* toolkit [57]. We did not remove hashtags and mentions because they can be informative for our study. We lowercased and tokenized using the *Spacy* toolkit [58]. Since the methods we used in this paper are all unsupervised, we did not split the data for training and testing (we will share our data according to Twitter policy once this paper has been accepted).

Topic Modeling

In order to analyze the public perception of long COVID, we applied CTM. CTM uses pretrained representations of language (eg, bidirectional encoder representations from transformers) to support topic modeling based on the context. To identify the best number of topics that can be extracted from the tweets with the best coherence and diversity between the topics, we applied the following automatic evaluation metrics with different settings: for coherence, we used normalized pointwise mutual information [59], Cv [60], and word embedding [61], while for diversity, we used topic diversity and inverted rank-biased overlap [62]. These metrics evaluate the semantic consistency and distinctiveness of topics, ensuring meaningful results.

First, we applied the measures on 5, 10, 15, and 20 topics. However, the best topic coherence and diversity metrics results were obtained by extracting 5 topics. We tried the same metrics over 4 and 6 topics to see whether they would have better results, but they did not. Moreover, to confirm the automatic evaluation we have shown the 5 extracted topics to the health care professionals.

To assess changes in topics of discussion over time, we compared timelines of topic distributions across the 3 regions and examined how this reflected change in public perception of long COVID. To discover topics and track the topic change over time, we constructed topic models on our Twitter data using zero-shot CTM implementation.

We used the *pyLDAvis* tool [63] to visualize the 5 topics generated by the CTM model. *PyLDAvis* visualizes the relationship between topics using multidimensional scaling. It helps understand topic overlap and distinctiveness, aiding in the interpretation of topic coherence and diversity metrics. This methodological detail enhances the reliability of our topic modeling results. Two public health experts examined the visualization and one of them labeled each topic to make it easier to understand the context of each topic. More specifically, we performed a basic analysis based on an examination of the estimates of the vector, a document-to-topic distribution, produced by the model. We first divided tweets into weekly buckets using Coordinated Universal Time–12 time stamps (eg, January 21-26, January 27 to February 2, and February 3-9, 2021). We then computed a mean vector for tweets in each bucket as done by Griffiths and Steyvers [64].

Sentiment Analysis

Sentiment analysis was conducted to gauge the emotional tone expressed in tweets related to long COVID. The objective was to understand how individuals on Twitter perceived and communicated their sentiments regarding long COVID throughout the year 2021. To analyze the sentiments expressed in tweets related to long COVID, we used sentiment analysis at the tweet level. For tweet sentiment classification, we used Llama 2, which have been recognized as one of the most widely used LLMs recently. The robustness of Llama 2 for sentiment analysis is supported by recent studies demonstrating its effectiveness across various domains [65-68].

The accuracy of sentiment analysis using Llama 2 was validated through manual inspection of a random sample of classified tweets. Misclassification rates were low, demonstrating the model's robustness. Prompts were designed to capture the overall sentiment of each tweet accurately, considering context and phrasing nuances.

The sentiment analysis pipeline provided by Hugging Face's Transformers library is designed to analyze the sentiment of a given piece of text. Pipelines are high-level objects that allow users to easily apply pretrained models to various natural language processing (NLP) tasks. This specific pipeline uses the Llama 2 model for sentiment analysis, facilitating efficient processing of text data. Through tokenization, encoding, and model inference, the pipeline predicts sentiment, categorizing it as positive, negative, or neutral. This process offers valuable insights into the emotional tone of tweets related to long COVID.

Ethical Considerations

This study was conducted using publicly available Twitter data, which are anonymized and deidentified. The data collection and analysis were carried out in accordance with Twitter's terms of service. No personal or sensitive information was accessed, ensuring the privacy and confidentiality of Twitter users.

As the research involved publicly available, anonymized data from Twitter, it did not require formal ethics review approval. This approach is consistent with TCPS 2 (Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans), article 2.2, regarding secondary analysis of public data where no direct interaction with human subjects occurs.

Given that only publicly accessible data were analyzed, specific informed consent from Twitter users was not required. The original consent provided by Twitter users for public sharing of their data under Twitter's terms of service allows for secondary analysis without additional consent.

All data analyzed in this study were anonymized and deidentified. No identifying information was collected, and data were handled in ways that fully comply with ethical guidelines on protecting participant confidentiality.

No compensation was provided to Twitter users, as the study used only publicly available data and did not involve direct participation or interaction with users.

Results

Topic Modeling

The 5 topics identified through topic modeling and a domain expert effort were as follows:

1. Long COVID in people including children in the context of vaccination (T1)
2. Duration and suffering associated with long COVID (T2)
3. Persistent symptoms of long COVID (T3)
4. The need for research on long COVID treatment (T4)
5. Measuring long COVID symptoms (T5)

A set of sample tweets associated with each topic was then examined to qualitatively validate the topic labels. The sample tweets provided contextual information that could not be inferred from the salient terms alone. Table 3 highlights the topics and their labels.

Table 3. Labels for top 5 prevalent topics in Canada, the United States, and Europe relating to long COVID[a].

Topic number	Salient terms (top 10 in italics)	Label based on salient terms	Label based on sample tweets
T1	*“people,” “children,” “kids,” “covid,” “long,” “get,” “many,” “risk,” “vaccinated,” “deaths,”* “know,” “cases,” “vaccine,” “schools,” “death,” “even,” “also,” “still,” “die,” “amp,” “term,” “vaccines,” “young,” “school,” “delta,” “health,” “protect,” “masks,” “spread,” and “immunity”	Long COVID in people, including children, in the context of vaccination	More objective perspectives: research, monitoring, experience, and news items highlighting significant impact of symptoms of long COVID in adults and children, including potential benefits of vaccination in reducing long COVID risk
T2	*“day,” “feel,” “like,” “back,” “days,” “year,” “got,” “time,” “pain,” “last,”* “week,” “feeling,” “months,” “since,” “one,” “better,” “friend,” “ago,” “still,” “felt,” “work,” “went,” “hope,” “years,” “body,” “go,” “march,” “going,” “first,” and “today”	Duration and suffering associated with long COVID	More subjective perspectives: suffering and frustration associated with long COVID
T3	*“symptoms,” “19,” “covid,” “long,” “study,” “post,” “patients,” “term,” “infection,” “haulers,”* “new,” “covid19,” “months,” “syndrome,” “coronavirus,” “fatigue,” “acute,” “effects,” “common,” “studies,” “viral,” “via,” “persistent,” “reported,” “sars,” “found,” “weeks,” “experiencing,” “brain,” and “lingering”	Persistent symptoms of long COVID	Persistent symptoms of long COVID, with greater focus on formal and informal advocacy and awareness-raising around long COVID
T4	*“longcovid,” “mecfs,” “research,” “please,” “cfs,” “support,” “amp,” “pwme,” “help,” “thank,”* “patients,” “community,” “need,” “share,” “us,” “longcovidkids,” “funding,” “sign,” “patient,” “awareness,” “join,” “millionsmissing,” “myalgicencephalomyelitis,” “treatments,” “illness,” “treatment,” “guidelines,” “nicecomms,” “needs,” and “thanks”	Need for research on long COVID treatment	Calls to address the plight of people with long COVID, either by researchers or by governments, or frustration at having been neglected, and frustration at having ongoing symptoms.
T5	*“fewer,” “ontario,” “vast,” “matters,” “measure,” “amongst,” “likelihood,” “possibility,” “polio,” “contract,”* “ha,” “tracking,” “victims,” “de,” “clinically,” “eg,” “vs,” “admitted,” “telegraph,” “exposure,” “estimate,” “fda,” “surge,” “impacted,” “reducing,” “70,” “american,” “ages,” “yo,” and “largely”	Measuring long COVID symptoms	The large number of people affected by, or projected to be affected, by long COVID symptoms, and associated societal impacts.

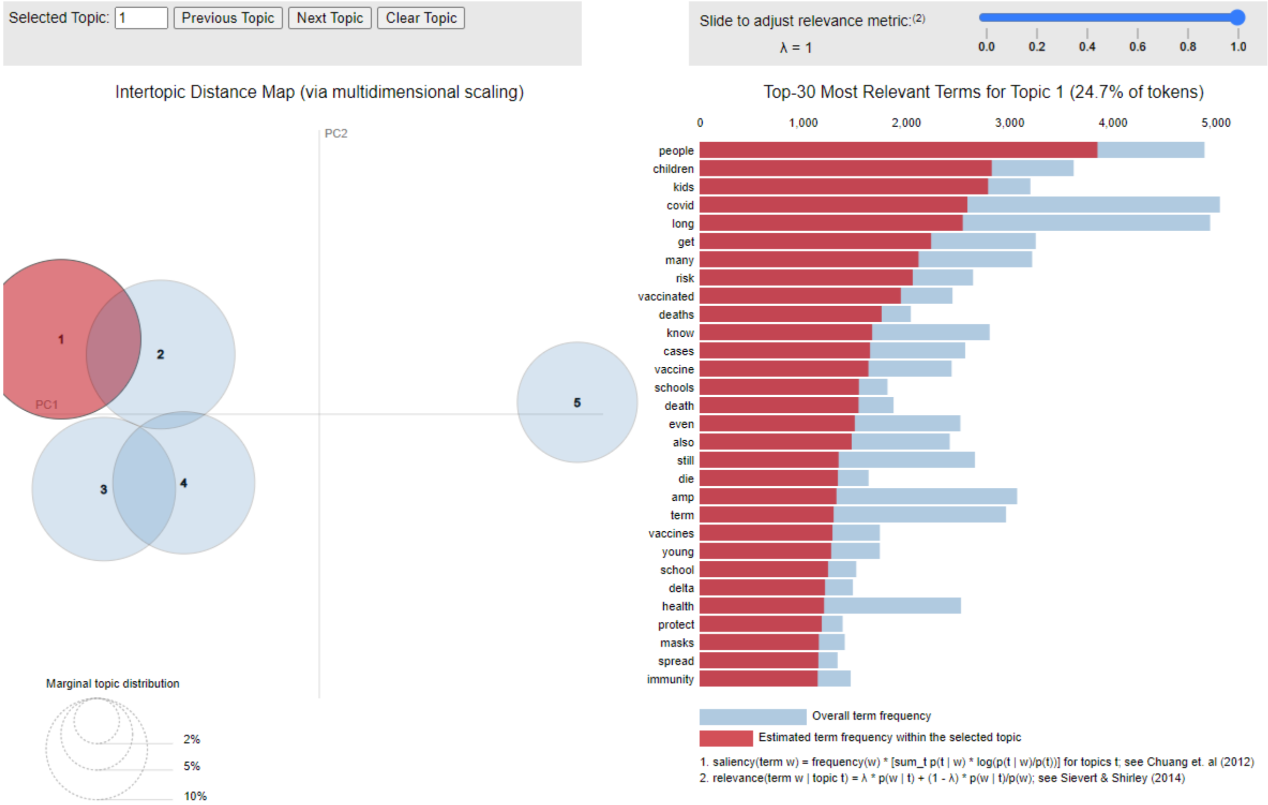
[a]Derived from topic modeling of Twitter content from 2021, with labels for each topic based on human interpretation of keywords and associated sample tweets. This table shows the topics identified, the salient terms associated with each topic, and the final labels given to each topic based on the interpretation of sample tweets.

Through expert's evaluation, it became clear that the accurate and meaningful labels could not be consistently developed through salient terms alone, devoid of the context of the sample tweets from which they were derived. In particular, we observed that topic T1 tended to be associated with more objective perspectives on long COVID such as research, monitoring, objective descriptions of experiences, and news items, whereas T2 tended to be associated with more subjective perspectives such as suffering and frustration associated with having long COVID. In addition, the sample tweets suggested that T3 and T4 had advocacy for more to be done to address long COVID as a prominent element. Finally, T5 had societal impacts as a prominent element within the sample tweets.

Examination of sample tweets by the health care professionals also revealed overlap in the content between some of the topics.

This is also reflected to some extent in the intertopic distances that were computed and visualized in Figure 1, which demonstrates some intertopic overlap among topics T1-T4 but not for T5. Since *pyLDAvis* is an interactive visualization of the topics, we asked the health care professionals to explore the adjacent metric settings. By default, *pyLDAvis* is set for λ=1, which sorts words just by their frequency within the specific topic (by their red bars). By contrast, setting λ=0 words sorts words by their "lift." This means that words whose red bars are nearly as long as their blue bars will be sorted at the top.

Figure 1. Intertopic distance map via multidimensional scaling. This figure visualizes the distance between the 5 topics identified from long COVID–related tweets in 2021, showing intertopic overlap among topics T1-T4 but not for T5. The study includes tweets from Canada, the United States, and Europe. T1: Long COVID in people including children in the context of vaccination; T2: duration and suffering associated with long COVID; T3: persistent symptoms of long COVID; T4: need for research on long COVID treatment; and T5: measuring long COVID symptoms. In the figure, T1 is selected.



Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

The types of discrepancies found between the salient terms–based topic labels and some of the sample tweets were due to low-label specificity, misrepresentation, or topic overlap. Some tweets were also found to be irrelevant to long COVID altogether; in particular, a portion of the sample tweets was irrelevant because of misclassification of long COVID–related tweets with tweets on COVID-19. Examples of these discrepancies are shown in Multimedia Appendix 1.

Based on the mean vector for each bucket, we drew graphs of long COVID topics over time as shown in Figures 2-7. We observed similar patterns between tweets in Canada and the United States. For example, the topic about subjective experiences around the duration and suffering associated with long COVID (T2) was less prominent early in the year but increased in prominence over time, peaking in late July-early August, and remained higher for the rest of the year, while there was no discernible trend in Europe. On the other hand, the topic on more objective perspectives on long COVID (T1) was relatively low in prominence in Europe compared with Canada and the United States. In contrast, the topic on persistent symptoms of long COVID (T3) was prominent throughout the year in all 3 regions.

Figure 2. Analyzing topic trends: visualizing the prominence of different topics over time through a bar chart for the Canada region. This figure shows the weekly distribution of the 5 identified topics in long COVID–related tweets in Canada during 2021.

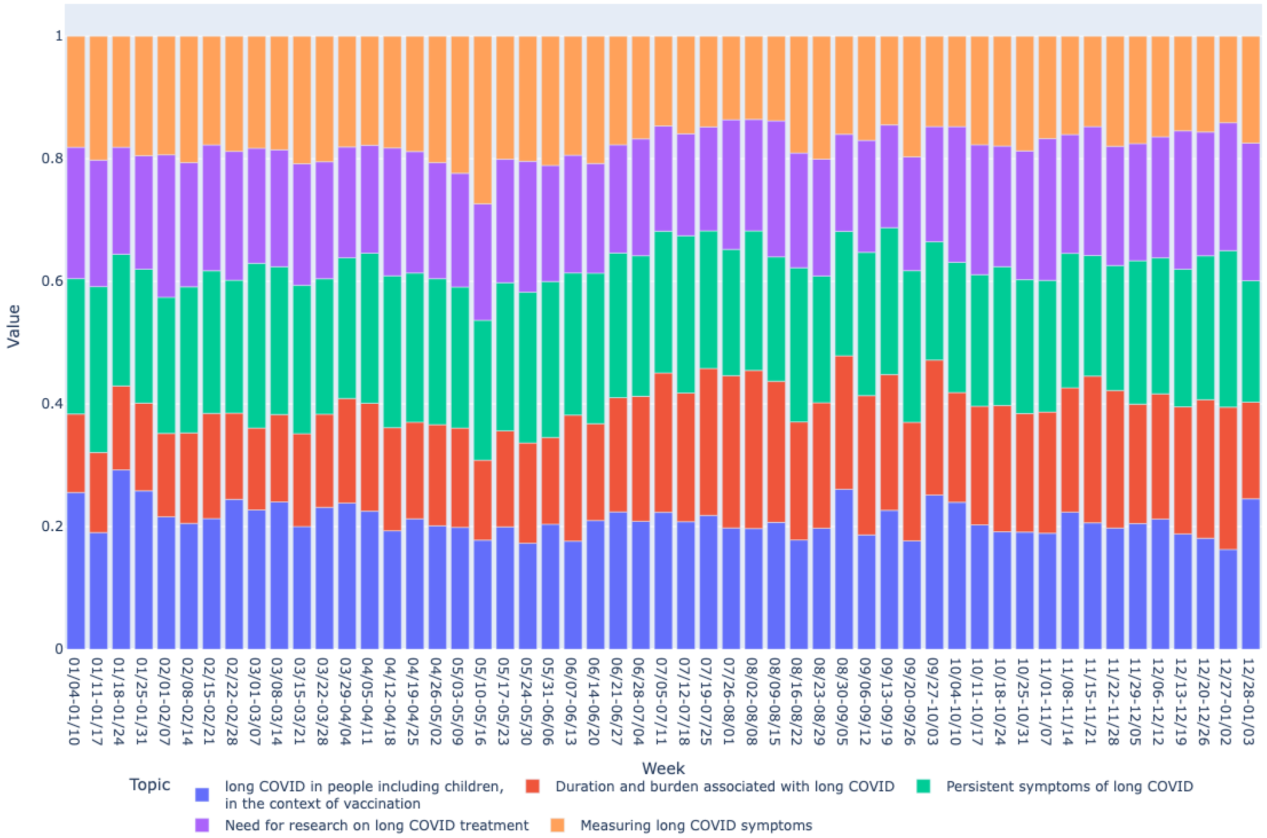


Figure 3. Analyzing topic trends: visualizing the prominence of different topics over time through a bar chart for the US region. This figure displays the weekly distribution of the 5 identified topics in long COVID–related tweets in the United States during 2021.

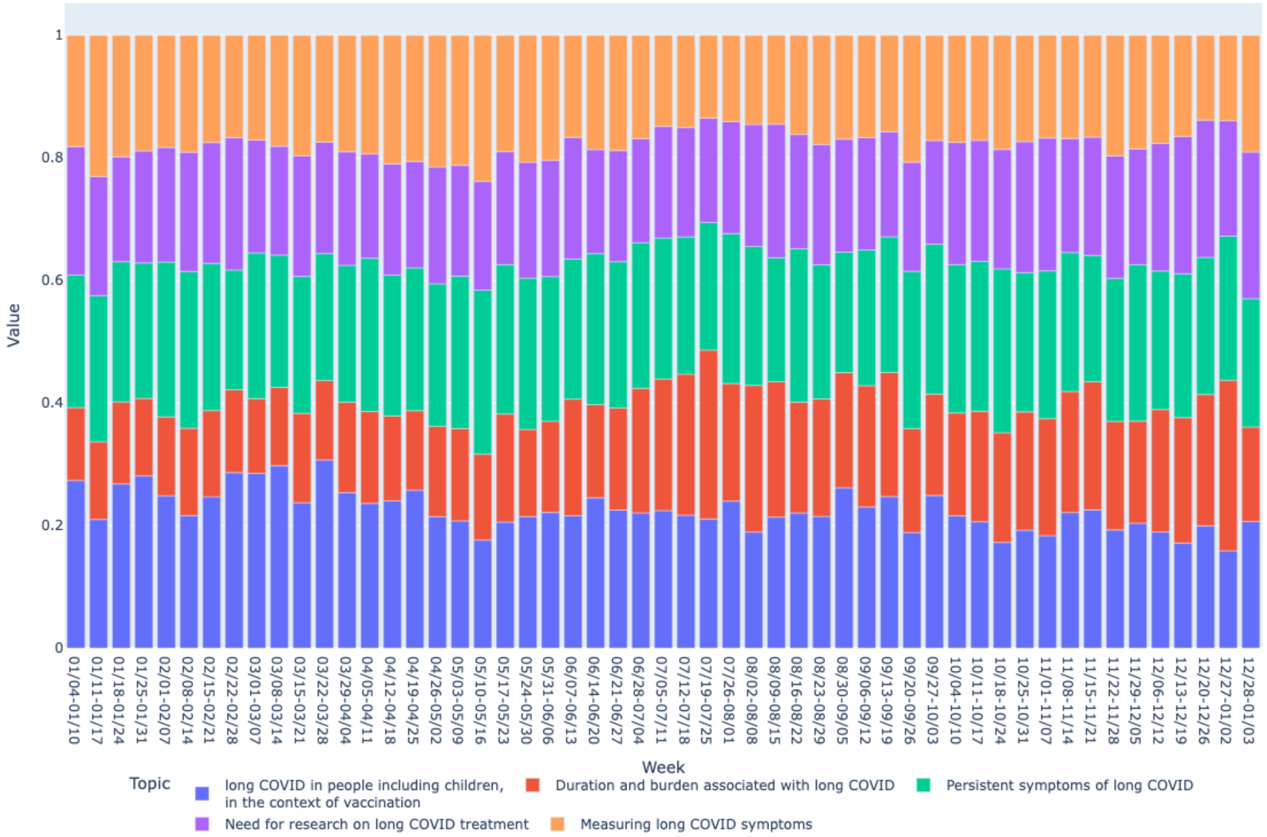


Figure 4. Analyzing topic trends: visualizing the prominence of different topics over time through a bar chart for Europe region. This figure depicts the weekly distribution of the 5 identified topics in long COVID–related tweets in Europe during 2021.

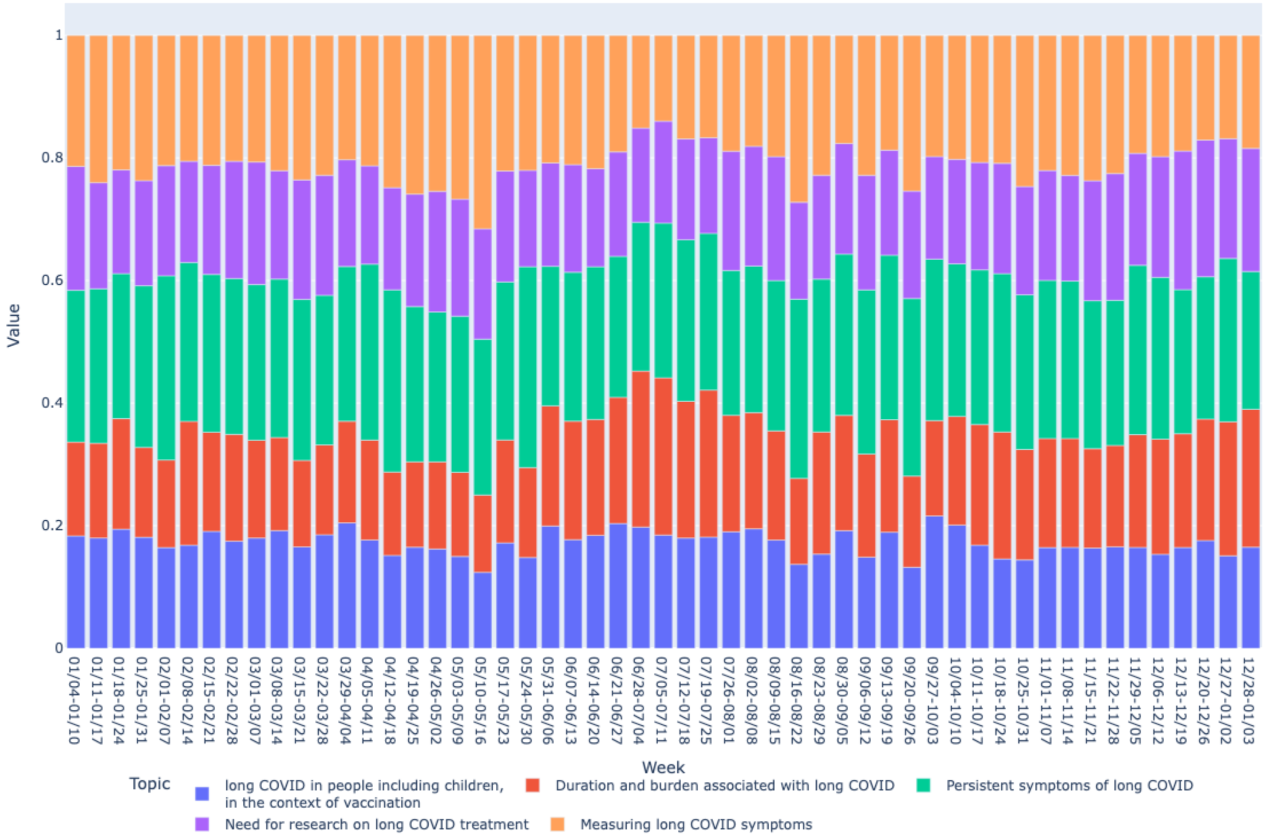


Figure 5. Analyzing topic trends: visualizing the prominence of different topics over time through a line chart for the Canada region. This figure shows the trends of the 5 identified topics in long COVID–related tweets in Canada over the year 2021.

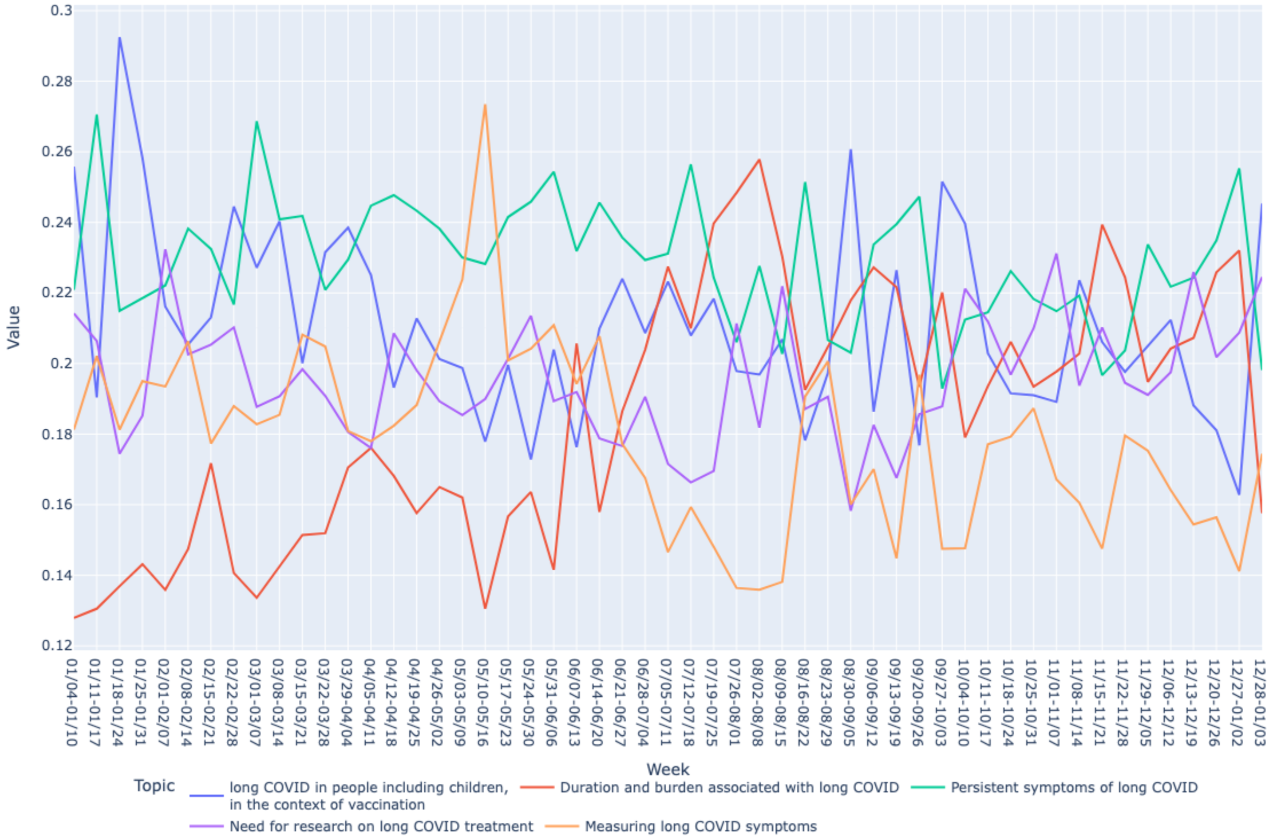


Figure 6. Analyzing topic trends: visualizing the prominence of different topics over time through a line chart for the US region. This figure illustrates the trends of the 5 identified topics in long COVID–related tweets in the United States over the year 2021.

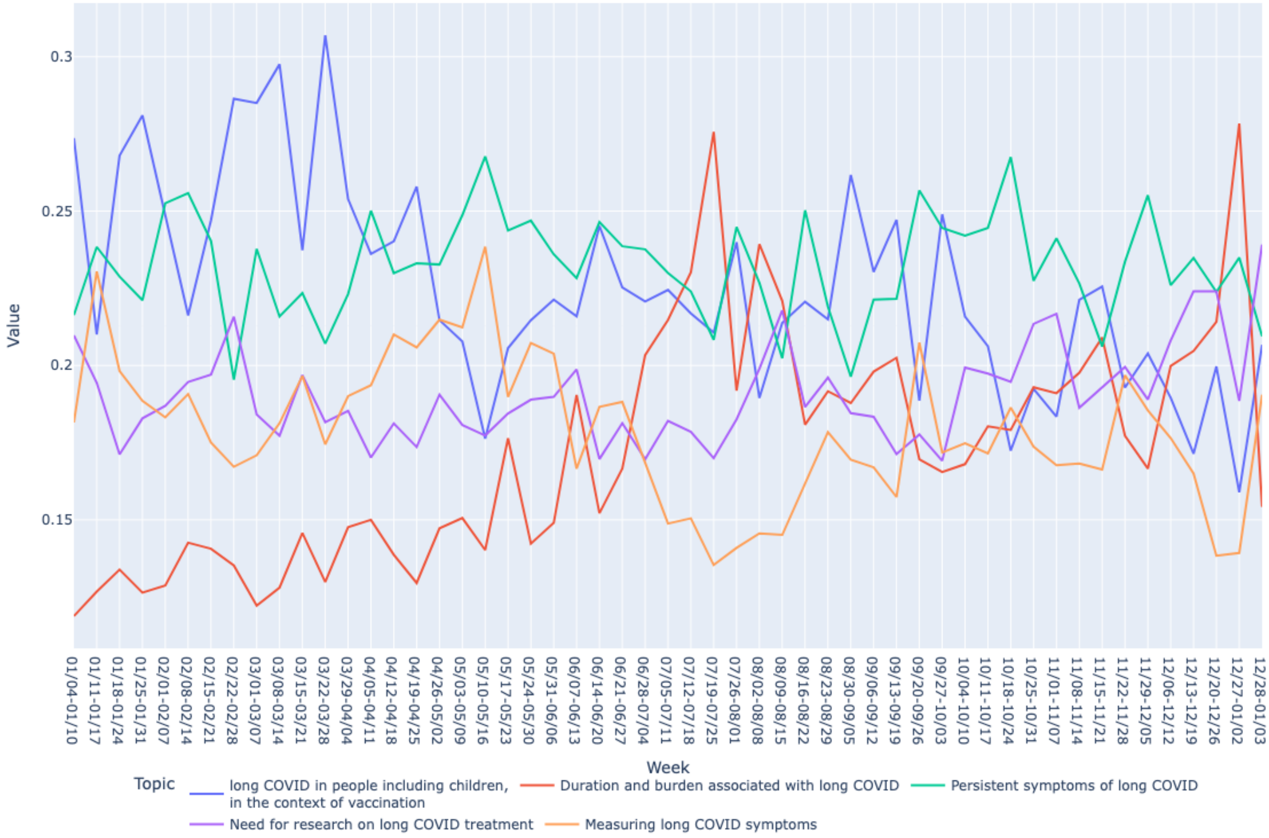
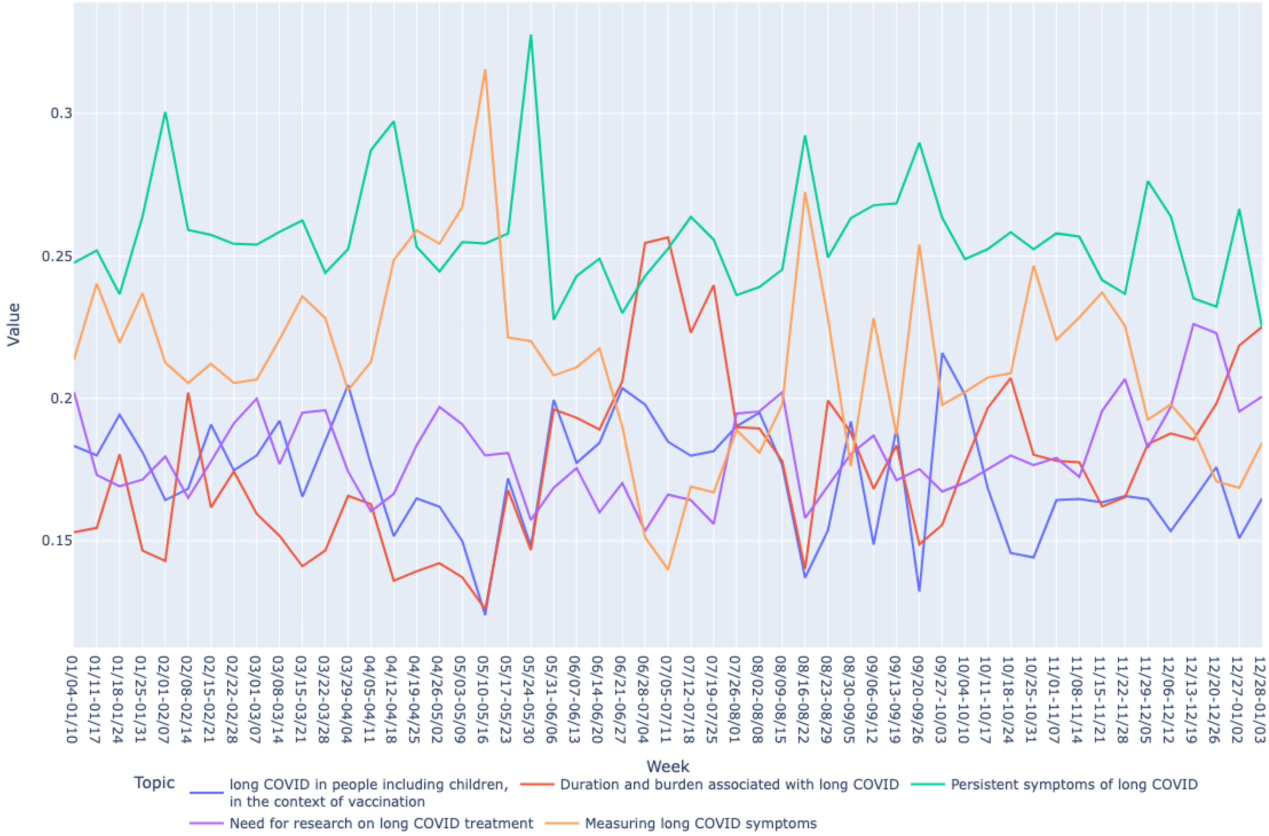


Figure 7. Analyzing topic trends: visualizing the prominence of different topics over time through a line chart for Europe region. This figure shows the trends of the 5 identified topics in long COVID–related tweets in Europe over the year 2021.



Both bar and line graphs were used to provide different perspectives on the data. Bar graphs offer a snapshot of weekly topic distribution, while line graphs illustrate trends over time, providing a comprehensive understanding of topic dynamics.

Sentiment Analysis

The sentiment distribution across long COVID–related tweets shows that there was no specific pattern. However, there was an increase in negative sentiment in July and August, followed by an increase in positive sentiment in September and again an increase in negative sentiment in December. In the United States, the trend was different, with a large amount of negative sentiment from January to March, a spike in July, and continued negative sentiment during August and September. April was notable for more positive sentiment and low negative sentiment. In particular, July had a lot of activity with some positive sentiment in addition to a large number of negative-sentiment tweets. The pattern in Europe was also different, with a large number of positive-sentiment tweets and substantial negative-sentiment tweets throughout the year.

There are many plausible explanations for these swings in positive and negative sentiments. For example, in July 2021, US President Joe Biden said that the long-term effects of COVID-19 can be considered a disability under federal civil rights laws, which may have spurred some positive sentiments [69]. In July, the National Institutes of Health also announced US $40 million for the study of long COVID and multisystem inflammatory syndrome in children [70]. In addition, multiple studies presented both positive and negative news about long COVID. All of these may have spurred both negative and positive discourse related to long COVID.

While we cannot establish a direct causal relationship, potential influences on sentiment swings can be hypothesized. Significant events, such as US President Joe Biden’s statement in July 2021 that the long-term effects of COVID-19 could be considered a disability under federal civil rights laws, and the National Institutes of Health’s announcement of US $40 million for the study of long COVID and multisystem inflammatory syndrome in children, likely contributed to these fluctuations. In addition, the presentation of both positive and negative news about long COVID in various studies may have further influenced public sentiment. These factors illustrate the complexity of inferring causality from observational data and highlight the multifaceted nature of public discourse.

Figures 8-10 show the count distributions of positive, negative, and neutral sentiments for the Canada, US, and European regions, respectively, providing a detailed view of the monthly sentiment counts. These figures help identify specific periods of increased positive or negative sentiments, which can be correlated with significant events.

Figures 11-13 show the percentage distributions of positive, negative, and neutral sentiments for the same regions, providing additional insights into the proportionate sentiment trends over the months. These figures complement the count-based visualizations by showing the relative proportions of each sentiment category, offering a more nuanced understanding of the sentiment landscape.

Figure 8. Sentiment analysis: visualizing the monthly sentiment counts produced by Llama 2 through a bar chart for the Canada region. This figure displays the count distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Canada during each month of 2021.

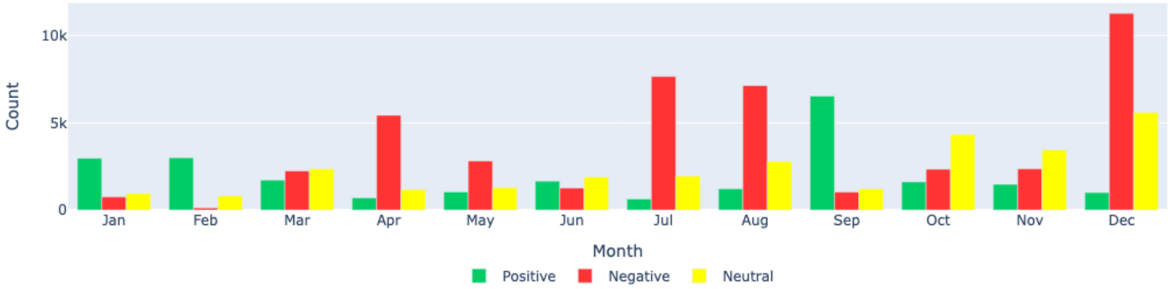


Figure 9. Sentiment analysis: visualizing the monthly sentiment counts produced by Llama 2 through a bar chart for the US region. This figure shows the count distribution of positive, negative, and neutral sentiments in long COVID–related tweets in the United States during each month of 2021.

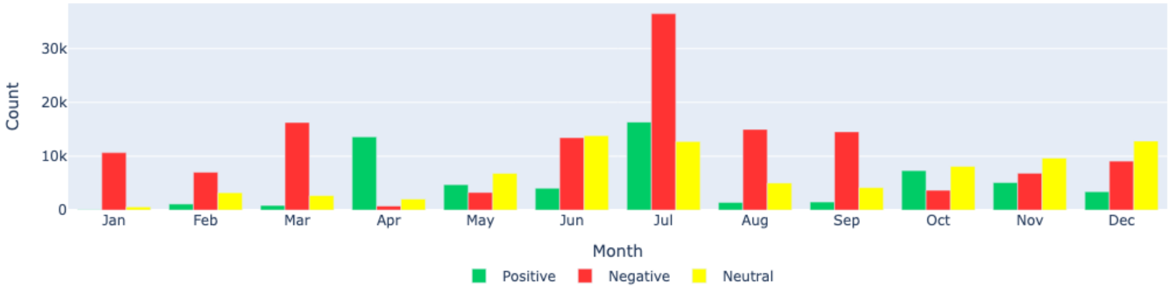


Figure 10. Sentiment analysis: visualizing the monthly sentiment counts produced by Llama 2 through a bar chart for Europe region. This figure illustrates the count distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Europe during each month of 2021.

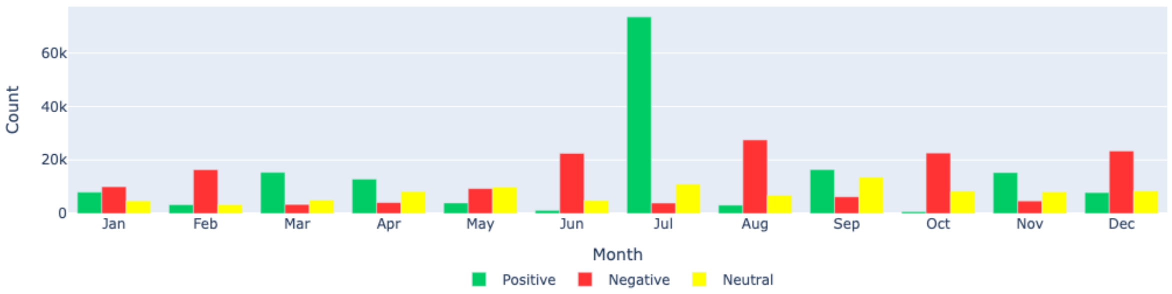


Figure 11. Sentiment analysis: visualizing the monthly sentiment percentages produced by Llama 2 through a bar chart for the Canada region. This figure displays the percentage distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Canada during each month of 2021.

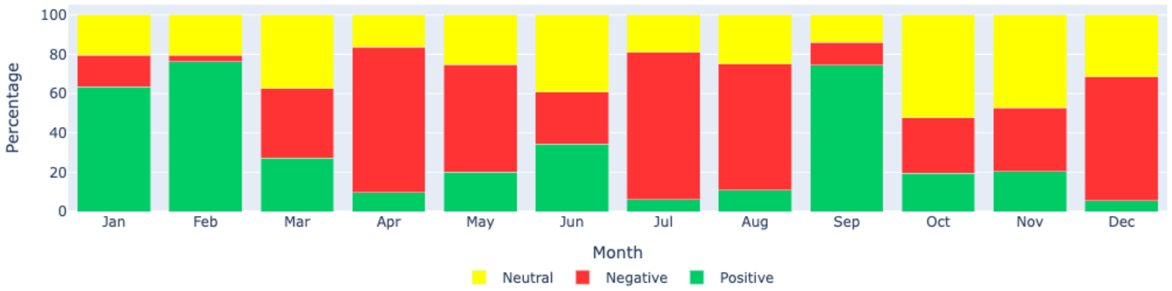


Figure 12. Sentiment analysis: visualizing the monthly sentiment percentages produced by Llama 2 through a bar chart for the US region. This figure shows the percentage distribution of positive, negative, and neutral sentiments in long COVID–related tweets in the United States during each month of 2021.

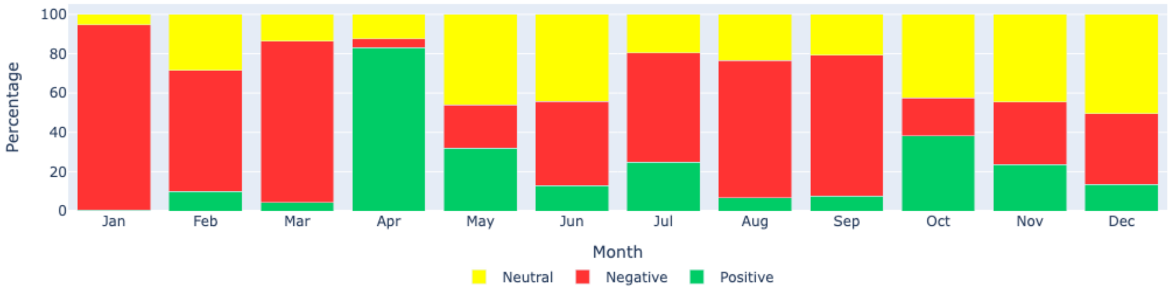
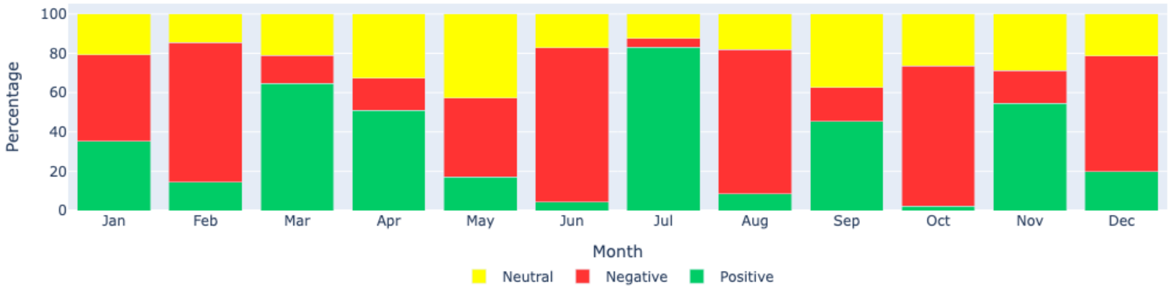


Figure 13. Sentiment analysis: visualizing the monthly sentiment percentages produced by Llama 2 through a bar chart for Europe region. This figure illustrates the percentage distribution of positive, negative, and neutral sentiments in long COVID–related tweets in Europe during each month of 2021.



Discussion

Principal Findings

In this study, we used topic modeling and sentiment analysis on Twitter data from Canada, the United States, and Europe in 2021. We identified 5 key topics: objective and subjective descriptions of the experience of people living with long COVID; persistent symptoms associated with long COVID; the need for more attention to be given to long COVID by researchers and governments; and measuring the impact of long COVID. The overlapping nature of these topics is likely due to the complex and ill-defined nature of long COVID, which shares symptoms with other conditions. Sentiment analysis showed a mix of positive and negative sentiments over time, with notable variations across different countries. The fluctuations in sentiment may be influenced by various events and news during the study period. For instance, misinformation such as claims that long COVID is merely psychological and not a real medical condition could lead to negative sentiments and confusion among the public. These insights into discussions and sentiments related to long COVID can help understand patient concerns and combat misinformation. The code used for data analysis is available in the study by AbuRaed [71].

The combined use of topic modeling and sentiment analysis in this study provided a richer, multidimensional view of the public discourse on long COVID. Topic modeling identified key themes such as persistent symptoms (T3) and the need for more research on treatments (T4), while sentiment analysis added an emotional layer to these discussions. For example, negative sentiments were consistently associated with topics about suffering and the call for more research, reflecting public frustration with the ongoing impact of long COVID. In contrast, topics such as vaccination (T1) often displayed more balanced or positive sentiments, indicating public optimism during vaccine rollout periods. By integrating these 2 approaches, we captured not only what people were discussing but also how they felt about these issues, offering a more comprehensive understanding of public sentiment and its fluctuations over time. This insight is crucial for public health communication and intervention strategies, as it highlights the emotional responses tied to different aspects of the long COVID discourse.

Most previous studies focused on either topic modeling or sentiment analysis related to long COVID but not both. In addition, many of these studies used LDA for topic modeling, which operates under the bag-of-words assumption and ignores word order. We used CTM, which captures word meanings in context, allowing for more coherent and semantically meaningful topics. For sentiment analysis, we used Llama 2, instead of rule-based or hybrid systems, demonstrating its robustness in sentiment classification. Our findings showed both consistencies and discrepancies compared with existing studies, underscoring the robustness of our identified topics and sentiment trends while also highlighting the nuanced capabilities of our methodology.

CTM addresses the limitations of traditional methods by capturing word meanings in context, enhancing topic coherence and relevance. This approach is particularly suited for complex and nuanced topics such as long COVID.

The topic labels, based on salient terms, were found to be inadequate upon examination of sample tweets. This discrepancy is expected due to the nonrepresentative nature of sample tweets and the overlapping representation among topics T1-T4, as indicated by the intertopic distances. Misinterpretation due to word-sense disambiguation challenges also contributed to misrepresentative labeling. Salient terms such as “like” or “long” are difficult to interpret without context, leading to irrelevance in the labels concerning long COVID.

The evaluation of topic labels involved expert review of sample tweets, ensuring that labels accurately reflected tweet content. This iterative process improves label accuracy and relevance. The methodology and results of this evaluation are detailed in the “Methods” and “Results” sections.

When examining topic prominence over time, it was challenging to identify clear patterns due to the overlapping nature of topics. However, there was an overall increasing trend in discussions around the duration and suffering associated with long COVID, consistent with the growing awareness and burden of long COVID as the pandemic progressed.

Public Health Implications

The results of this study have significant public health implications. Monitoring discussions and perceptions of people with long COVID can provide valuable insights into their needs and changing issues over time. Although geographical analysis is limited due to data constraints, this system can assess the impact of various actions and measures across jurisdictions. Furthermore, this approach can be applied to other emerging public health issues to monitor discourse, address concerns, and correct misinformation.

By identifying key topics and sentiment trends, our study can inform public health messaging and interventions to address misinformation. For example, if we observe a spike in negative sentiment and misinformation regarding the efficacy of long COVID treatments, public health officials can launch targeted information campaigns to provide accurate information and resources about available treatments and their effectiveness. Understanding public concerns and sentiment shifts allows for these targeted information campaigns to correct false narratives and provide timely support to affected individuals.

Limitations

There were several limitations to this study related to the data source, data collection and processing, the NLP methods used, and the approach to interpretation of the outputs.

With regard to the data source, a portion of the tweets were cut off, either because Twitter’s character limit per post forced the user to post multiple tweets to express their thoughts or because of the extraction process itself. In addition, some tweets are repeated multiple times (sometimes more than 80 times), because of retweets. Moreover, because of the design of the Twitter platform, such as character limits, and the intentions of Twitter users (eg, to grab people’s attention), the ideas expressed are abbreviated and unclear. We extracted only those tweets that were in the English language and further limited tweets to those originating from Canada, the United States, and Europe. Thus, the data are unlikely to be representative of discussions and perceptions on long COVID elsewhere in the world. Finally, Twitter users constitute a small subset of the world population, further limiting the representativeness of these data.

With regard to data collection and data processing, the collection of Twitter data related to long COVID using long COVID–related keywords was subject to misclassification, since tweets relating to long COVID had significant overlap with tweets unrelated to long COVID. While this misclassification cannot be quantified without examining and categorizing each tweet one by one, an examination of a random sample of 164 tweets identified that approximately 13% (21/164) of the tweets were unrelated to long COVID. This misclassification may be in part because of the still poorly defined nature of long COVID as a condition, and in part a result of the long COVID–related keywords that were used to extract the tweets. For example, fibromyalgia, a chronic condition with similar symptomatology to long COVID, was discussed in some tweets that were irrelevant to long COVID.

With respect to NLP methods used to analyze the Twitter data, there was significant overlap between many of the topics in the topic model, as discussed above. In addition, challenges with word-sense disambiguation limit the interpretability of results.

With respect to the approach to interpretation of the outputs, having 1 domain expert to label the topics based on the sample tweets and a second domain expert to label the topics based on the salient terms would have helped enhance objectiveness in the labeling process. In this study, since the same domain expert did the labeling using these 2 approaches, there was likely some bias in how the labels were determined. Moreover, while no formal method was applied in this study to account for how public health measures impacted public perceptions on long COVID, the change in public health measures over time and region over the course of 2021 does constitute a potential source of bias and confounding in our analysis. For example, more stringent measures may prompt a more favorable perception toward government among those concerned about long COVID, since stringent measures may be perceived as part of an effort to reduce the risk to individuals not only of COVID-19 but also of long COVID. Conversely, less stringent measures may be perceived less favorably for similar reasons.

Conclusions

Our key findings include the identification of 5 main topics related to long COVID, significant regional variations in public discourse, and the fluctuating nature of sentiments influenced by major events. These findings provide a nuanced understanding of public concerns and the potential to inform public health strategies.

This study shows exploratory results from topic modeling and sentiment analysis on long COVID–related tweets in Canada, the United States, and Europe. By comparing results across these regions, we demonstrated changes in topic prominence over time. Public health domain experts played a crucial role in interpreting the results, emphasizing the importance of a human-in-the-loop approach in sentiment analysis. Although we identified some regional and temporal differences in topics, the main interpretation is the evolving public discourse and increasing awareness of long COVID throughout 2021.

The insights gained from this study can help public health officials and policy makers design more effective communication strategies and interventions tailored to the needs and concerns of those experiencing long COVID. By understanding the evolving public sentiment and key issues discussed on social media, public health responses can be better aligned with the population’s needs. This research contributes to the broader understanding of how social media data can be leveraged for public health surveillance and intervention.

Future Directions

Future studies should explore more refined methods to enhance the specificity and accuracy of topic modeling and sentiment analysis. In addition, integrating more diverse social media platforms and languages could provide a more comprehensive understanding of global long COVID discourse. Finally, continuous monitoring and real-time analysis of social media data can further aid in understanding public perceptions and developing timely public health interventions.

Acknowledgments

The authors acknowledge the Canadian Institutes of Health Research (CIHR) and Michael Smith for Health System Impact Fellowship. The author thank Gaetano Cimino for providing feedback on the paper. This manuscript was prepared with the assistance of an artificial intelligence–based tool for language processing and editing, specifically ChatGPT by OpenAI [72], which was used to improve language clarity and coherence. All content and conclusions are the responsibility of the authors.

Conflicts of Interest

NZJ participated in advisory boards and has spoken for AbbVie and Gilead, not related to this work.

Multimedia Appendix 1

Sample tweets.
[DOCX File , 15 KB-Multimedia Appendix 1]

References

1. Coronavirus (COVID-19) dashboard. World Health Organization. URL: https://covid19.who.int [accessed 2024-06-30]
2. Altmann DM, Whettlock EM, Liu S, Arachchillage DJ, Boyton RJ. The immunology of long COVID. Nat Rev Immunol. 2023;23(10):618-634. [doi: 10.1038/s41577-023-00904-7] [Medline: 37433988]
3. Binka M, Klaver B, Cua G, Wong AW, Fibke C, Velásquez García HA, et al. An elastic net regression model for identifying long COVID patients using health administrative data: a population-based study. Open Forum Infect Dis. 2022;9(12):ofac640. [doi: 10.1093/ofid/ofac640] [Medline: 36570972]
4. Long-term effects of COVID-19. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html#:~:text=People%20with%20Long%20COVID%20can,can%20sometimes%20result%20in%20disability [accessed 2024-06-30]
5. Natarajan A, Shetty A, Delanerolle G, Zeng Y, Zhang Y, Raymont V, et al. A systematic review and meta-analysis of long COVID symptoms. Syst Rev. 2023;12(1):88. [FREE Full text] [doi: 10.1186/s13643-023-02250-0] [Medline: 37245047]
6. Zhang H, Zang C, Xu Z, Zhang Y, Xu J, Bian J, et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. Nat Med. 2023;29(1):226-235. [FREE Full text] [doi: 10.1038/s41591-022-02116-3] [Medline: 36456834]
7. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. J Med Internet Res. 2020;22(4):e19118. [doi: 10.2196/19118] [Medline: 32302966]
8. Liu J, Nie H, Li S, Chen X, Cao H, Ren J, et al. Tracing the pace of COVID-19 research: topic modeling and evolution. Big Data Res. 2021;25:100236. [doi: 10.1016/j.bdr.2021.100236]
9. Dong M, Cao X, Liang M, Li L, Liu G, Liang HY. Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling. medRXiv. Preprint posted online April 10. [doi: 10.1101/2020.03.26.20044164]
10. Sussman KL, Bouchacourt L, Bright LF, Wilcox GB, Mackert M, Norwood AS, et al. COVID-19 topics and emotional frames in vaccine hesitation: a social media text and sentiment analysis. Digit Health. 2023;9:20552076231158308. [FREE Full text] [doi: 10.1177/20552076231158308] [Medline: 36896330]
11. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. J Gen Intern Med. 2020;35(7):2244-2247. [doi: 10.1007/s11606-020-05889-w] [Medline: 32399912]
12. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with twitter data in the United States: spatiotemporal perspective. J Med Internet Res. 2021;23(9):e30854. [FREE Full text] [doi: 10.2196/30854] [Medline: 34346888]
13. Ridhwan MK, Hargreaves CA. Leveraging twitter data to understand public sentiment for the COVID‐19 outbreak in Singapore. Int J Inf Manag Data Insights. 2021;1(2):100021. [doi: 10.1016/j.jjimei.2021.100021]
14. Huangfu L, Mo Y, Zhang P, Zeng DD, He S. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling. J Med Internet Res. 2022;24(2):e31726. [FREE Full text] [doi: 10.2196/31726] [Medline: 34783665]
15. Beliga S, Martinčić-Ipšić S, Matešić M, Petrijevčanin Vuksanović I, Meštrović A. Infoveillance of the Croatian online media during the COVID-19 pandemic: one-year longitudinal study using natural language processing. JMIR Public Health Surveill. 2021;7(12):e31540. [FREE Full text] [doi: 10.2196/31540] [Medline: 34739388]
16. Jafarzadeh H, Pauleen DJ, Abedin E, Weerasinghe K, Taskin N, Coskun M. Making sense of COVID-19 over time in New Zealand: assessing the public conversation using Twitter. PLoS One. 2021;16(12):e0259882. [FREE Full text] [doi: 10.1371/journal.pone.0259882] [Medline: 34910732]
17. Maulana FI, Adi PDP, Lestari D, Purnomo A, Prihatin SY. Twitter data sentiment analysis of COVID-19 vaccination using machine learning. 2022. Presented at: 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI); December 8-9, 2017:582-587; Yogyakarta, Indonesia. [doi: 10.1109/ISRITI56927.2022.10053035]
18. Marcec R, Likic R. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgrad Med J. 2022;98(1161):544-550. [doi: 10.1136/postgradmedj-2021-140685] [Medline: 34373343]
19. Kumar AHS, Shausan A, Demartini G, Rahimi A. Automatic identification of 5C vaccine behaviour on social media. 2022. Presented at: Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022); October 10, 2024:136-146; Gyeongju, Republic of Korea.
20. Akpatsa SK, Addo PC, Lei H, Li X, Dorgbefu M, Fiawoo DD, et al. Sentiment analysis and topic modeling of Twitter data: a text mining approach to the US-Afghan war crisis. SSRN Journal. Preprint posted online on March 23, 2022. [doi: 10.2139/ssrn.4064560]
21. Guo Y, Zhu J, Huang Y, He L, He C, Li C, et al. Public opinions toward COVID-19 vaccine mandates: a machine learning-based analysis of U.S. tweets. AMIA Annu Symp Proc. 2022;2022:502-511. [FREE Full text] [Medline: 37128441]
22. Maulana FI, Heryadi Y, Suparta W, Arifin Y. Social media analysis using sentiment analysis on COVID-19 from twitter. 2022. Presented at: 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE); December 13-14, 2022:286-290; Yogyakarta, Indonesia. [doi: 10.1109/icitisee57756.2022.10057932]

23. Masood A, Iqbal MM, Nayab A, Farooq M, Saeed MS. Impact of COVID-19 on human health using social media sentiment analysis. J Comput Biomed Inf. 2023;5(01):41-51.
24. Upadhyay K. Topic modelling and sentiment analysis architecture for social networks in the COVID pandemic. Int J Comput Artif Intell. 2022;3(2):56-59. [doi: 10.33545/27076571.2022.v3.i2a.55]
25. Sha H, Al HM, Mohler G, Brantingham P. Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. arXiv. Preprint posted online on April 19, 2020. [FREE Full text]
26. Hosseini P, Hosseini P, Broniatowski DA. Content analysis of Persian/Farsi tweets during the COVID-19 pandemic in Iran using NLP. arXiv. Preprint posted online on May 17, 2020. [FREE Full text] [doi: 10.18653/v1/2020.nlpcovid19-2.26]
27. Sharma K, Seo S, Meng C, Rambhatla S, Liu Y. COVID-19 on social media: analyzing misinformation in Twitter conversations. arXiv. Preprint posted online on March 26, 2020. [FREE Full text]
28. Odlum M, Cho H, Broadwell P, Davis N, Patrao M, Schauer D, et al. Application of topic modeling to tweets as the foundation for health disparity research for COVID-19. Stud Health Technol Inform. Jun 26, 2020;272:24-27. [FREE Full text] [doi: 10.3233/SHTI200484] [Medline: 32604591]
29. Wang X, Zou C, Xie Z, Li D. Public opinions towards COVID-19 in California and New York on Twitter. medRxiv. Preprint posted online on July 14, 2020. [doi: 10.1101/2020.07.12.20151936] [Medline: 32699856]
30. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res. 2020;22(4):e19016. [doi: 10.2196/19016] [Medline: 32287039]
31. Ordun C, Purushotham S, Raff E. Exploratory analysis of Covid-19 tweets using topic modeling, UMAP, and DiGraphs. arXiv. Preprint posted online on May 6, 2020. [FREE Full text]
32. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res. Jan 2003;3:993-1022. [doi: 10.7551/mitpress/1120.003.0082]
33. Southwick L, Guntuku SC, Klinger EV, Pelullo A, McCalpin H, Merchant RM. The role of digital health technologies in COVID-19 surveillance and recovery: a specific case of long haulers. Int Rev Psychiatry. 2021;33(4):412-423. [doi: 10.1080/09540261.2020.1854195] [Medline: 33860736]
34. Fu YB. Investigating public perceptions regarding the long COVID on Twitter using sentiment analysis and topic modeling. Med Data Min. 2022;5(4):24. [FREE Full text] [doi: 10.53388/mdm20220520024]
35. Yin J, Wang J. A Dirichlet multinomial mixture model-based approach for short text clustering. 2014. Presented at: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14); August 24-27, 2014:233-242; New York, NY. [doi: 10.1145/2623330.2623715]
36. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, et al. Conditional random fields as recurrent neural networks. 2015. Presented at: Proceedings of the IEEE International Conference on Computer Vision; June 20-23, 1995:1529-1537; Cambridge, MA. [doi: 10.1109/iccv.2015.179]
37. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. 2013. Presented at: Proceedings of the 22nd International Conference on World Wide Web; 2013 May 13-17:1445-1456; Rio de Janeiro, Brazil. [doi: 10.1145/2488388.2488514]
38. Hutto CJ, Gilbert EE. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014. Presented at: Eighth International Conference on Weblogs and Social Media (ICWSM-14); May 16, 2014:216-225; Ann Arbor, MI. [doi: 10.1609/icwsm.v8i1.14550]
39. Nielsen FÅ. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. 2011. Presented at: Proceedings of the ESWC2011 Workshop on "Making Sense of Microposts": big things come in small packages; May 30, 2011:93-98; Crete, Greece. URL: http://arxiv.org/abs/1103.2903
40. An TK, Kim MH. A new diverse AdaBoost classifier. 2010. Presented at: Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence—Volume 01; October 23-24, 2010:359-363; Sanya, China. [doi: 10.1109/AICI.2010.82]
41. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016. Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
42. Haykin S. Neural Networks: A Comprehensive Foundation. Hoboken, NJ. Prentice Hall PTR; 1994.
43. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. In: Encyclopedia of Machine Learning. Boston, MA. Springer; 2010:713-714.
44. Suthaharan S, Suthaharan S. Support vector machine. In: Machine Learning Models and Algorithms for Big Data Classification: Thinking With Examples for Effective Learning. Boston, MA. Springer; 2016:207-235.
45. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on October 11, 2018. [FREE Full text]
46. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-Based sentiment analysis. J Med Internet Res. 2021;23(2):e25431. [doi: 10.2196/25431] [Medline: 33497352]
47. Déguilhem A, Malaab J, Talmatkadi M, Renner S, Foulquié P, Fagherazzi G, et al. Identifying profiles and symptoms of patients with long COVID in France: data mining infodemiology study based on social media. JMIR Infodemiology. 2022;2(2):e39849. [doi: 10.2196/39849] [Medline: 36447795]

48. Bhattacharyya A, Seth A, Rai S. The effects of long COVID-19, its severity, and the need for immediate attention: analysis of clinical trials and Twitter data. Front Big Data. 2022;5:1051386. [doi: 10.3389/fdata.2022.1051386] [Medline: 36588926]

49. Mohammad SM, Turney PD. NRC Emotion Lexicon. Ottawa, ON. National Research Council; 2013:234.

50. Bradley MM, Lang PJ. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Gainesville, FL. Technical Report C-1, the Center for Research in Psychophysiology, University of Florida; 1999:25-36.

51. Miyake E, Martin S. Long Covid: online patient narratives, public health communication and vaccine hesitancy. Digit Health. 2021;7:20552076211059649. [FREE Full text] [doi: 10.1177/20552076211059649] [Medline: 34868622]

52. Ramakrishnan K, Balakrishnan V, Han GJ, Seong NK. Long COVID emotion analyzer: using machine learning approach. 2023. Presented at: Proceedings of the 9th International Conference on Computational Science and Technology; August 27-28, 2022:529-541; Johor Bahru, Malaysia. [doi: 10.1007/978-981-19-8406-8_42]

53. Bianchi F, Terragni S, Hovy D, Nozza D, Fersini E. Cross-lingual contextualized topic models with zero-shot learning. arXiv. Preprint posted online on April 16, 2020. [FREE Full text]

54. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv. Preprint posted online on February 27, 2023. [FREE Full text]

55. Twitter Developer Platform. URL: https://developer.twitter.com [accessed 2024-06-30]

56. spacy-langdetect 0.1.2. The Python Package Index. 2020. URL: https://pypi.org/project/spacy-langdetect/ [accessed 2020-06-15]

57. tweet-preprocessor 0.6.0. The Python Package Index. 2020. URL: https://pypi.org/project/tweet-preprocessor/ [accessed 2020-06-15]

58. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Scientometrics Research. URL: https://sentometrics-research.com/publication/72/ [accessed 2024-10-10]

59. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. 2015. Presented at: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; February 2-6, 2015:399-408; Shanghai, China. [doi: 10.1145/2684822.2685324]

60. Syed S, Spruit M. Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. 2017. Presented at: IEEE International Conference on Data Science and Advanced Analytics (DSAA); October 19-21, 2017:165-174; Tokyo, Japan. [doi: 10.1109/DSAA.2017.61]

61. Fang A, Macdonald C, Ounis I, Habel P. Using word embedding to evaluate the coherence of topics from twitter data. 2016. Presented at: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval; July 17-21, 2016:1057-1060; Pisa, Italy. [doi: 10.1145/2911451.2914729]

62. Bianchi F, Terragni S, Hovy D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. 2020. Presented at: 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; August 1-6, 2021:759-766; Online. [doi: 10.18653/v1/2021.acl-short.96]

63. Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. 2014. Presented at: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces; June 27, 2014:63-70; Baltimore, MD. [doi: 10.3115/v1/W14-3110]

64. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A. 2004;101 Suppl 1(Suppl 1):5228-5235. [doi: 10.1073/pnas.0307752101] [Medline: 14872004]

65. Jeong C. Fine-tuning and utilization methods of domain-specific LLMs. arXiv. Preprint posted online on January 1, 2024. [FREE Full text]

66. Pavlyshenko BM. Analysis of disinformation and fake news detection using fine-tuned large language model. arXiv. Preprint posted online on September 9, 2023. [FREE Full text]

67. Yang H, Zhang Y, Xu J, Lu H, Heng PA, Lam W. Unveiling the generalization power of fine-tuned large language models. arXiv. Preprint posted online on April 14, 2024. [FREE Full text]

68. Pavlyshenko BM. Financial news analytics using fine-tuned Llama 2 GPT model. arXiv. Preprint posted online on August 24, 2023. [FREE Full text]

69. Biden administration recognizes long COVID as a disability under federal civil rights laws. The Washington Post. 2021. URL: https://www.washingtonpost.com/politics/biden-ada-long-covid-disability/2021/07/26/972f2a04-ee20-11eb-a452-4da5fe48582d_story.html [accessed 2024-06-30]

70. Children's National Hospital and NIAID launch study on long-term impacts of COVID-19 in kids. Children's National Hospital. 2021. URL: https://www.childrensnational.org/about-us/newsroom/2021/cnh-and-niaid-launch-large-study-on-long-term-impacts-of-covid-19-and-mis-c-on-kids [accessed 2024-06-30]

71. AbuRaed A. COVID-19 data and analysis. GitHub repository. URL: https://github.com/AhmedAbuRaed/covid19 [accessed 2024-06-30]

72. ChatGPT. OpenAI. URL: https://chatgpt.com/ [accessed 2024-12-03]

Abbreviations

CTM: contextualized topic modeling
LDA: latent Dirichlet allocation
LLM: large language model
NLP: natural language processing

Edited by A Mavragani; submitted 12.04.24; peer-reviewed by A Yagahara, N Yiannakoulias; comments to author 22.06.24; revised version received 28.07.24; accepted 24.09.24; published 09.12.24

Please cite as:
AbuRaed AGT, Prikryl EA, Carenini G, Janjua NZ
Long COVID Discourse in Canada, the United States, and Europe: Topic Modeling and Sentiment Analysis of Twitter Data
J Med Internet Res 2024;26:e59425
URL: https://www.jmir.org/2024/1/e59425
doi: 10.2196/59425
PMID: